Benchmarks Online

Skip Navigation Links

Page One

Campus Computing News

Spring Break Hours

EIS News

Today's Cartoon

RSS Matters

The Network Connection

Link of the Month

WWW@UNT.EDU

Short Courses

IRC News

Staff Activities

Subscribe to Benchmarks *Online*

Research and Statistical Support
University of North Texas

RSS Matters

Link to the last RSS article here: Resolving A Case of An Expired SAS 8.2 Installation. -- Ed.

Null Hypothesis Significance Testing

By Mike Clark, Research and Statistical Support Services Consultant

This is the beginning of a series discussing methodological approaches used in the social sciences. This article outlines the general problems and difficulties associated with a common method of statistical inference in psychology (my background) and other social science fields: null hypothesis significance testing (hereafter NHST). This introduction can serve as a starting point for researchers that are interested in examining these important issues in further detail. Subsequent articles will discuss alternative inference frameworks such as Bayesian analysis and Likelihood Estimation.

A Conceptual Overview

Statistical hypothesis testing involves setting up an initial hypothesis, and then performing a set of calculations on the data that give us some basis to judge as to whether our initial hypothesis should be retained or rejected. A common example in the social sciences is the situation where the researcher is interested in whether the means of various groups differ in a specified population. For example, we may want to see if grade point averages vary across college classification (freshman, sophomore, junior, senior). Following NHST procedures, a hypothesis that we might initially hold is that there is no difference among the groups (i.e. that their means are equal). We then perform our statistical analysis, and our procedures may lead us to say that we have not provided enough support to reject our initial hypothesis, or our procedures may lead us to believe that the initial hypothesis is untenable, whereby we would conclude that there are differences in the population groups. Subsequently, more statistical analyses, using similar logic, would be performed to discover specifically which groups differ.

The Problems

A thought that might occur to many researchers in a discussion of NHST is that they didn't know there was a problem or they may have been only vaguely aware of viable alternatives to NHST. This has been the case in many basic statistics courses – students are not told that there are some subtle difficulties with NHST, and that other alternatives might be more appropriate depending on what the researcher is trying to accomplish.

An important issue which is sometimes overlooked involves the practical interpretation of what we are doing. In NHST, we may state a null hypothesis that the difference between population groups are zero, or if we have more information, we may specify a specific value (or in a single sample case, we may specify that the mean of the sample data is equal to the population mean). Nonetheless, it is almost impossible to come up with that exact specified value in the sample under any circumstances of adequate sampling of data. For example, in a two-group design, our null hypothesis states that there are no population group differences while the alternative hypothesis states that the population groups are not equal. Below is an example of a more formal expression of hypotheses (null and alternative) regarding the difference between two group means:

$$\begin{gathered} H_0 : \mu_X - \mu_Y = 0 \\ H_A : Not \quad H_0 \end{gathered}$$

No matter how much the freshmen and sophomore populations look similar, the odds of them having *exactly the same sample GPA*, regardless of class sizes, is next to zero, and yet this is what our null hypothesis is suggesting. This null hypothesis of no difference in the population is sometimes thought of as a “straw-man” statement since we know that group samples will reflect some differences to some arbitrary decimal point. Having the observed sample difference, however small, be declared as statistically significant, is then a function of having a large enough sample size (all other things being equal) – if statistical significance is needed for the observed sample difference, one only needs to increase the sample sizes until the observed p-value reaches the cut-off criterion for significance.

Another source of confusion is related to the interpretation of NHST analysis results. Common sense would suggest that we are trying to determine the viability of a hypothesis. In other words, what the probability is that a hypothesis is true given the data at hand [p(H|D), the probability of an hypothesis given the data]. On the contrary, NHST actually involves a different conditional statement. We are not looking for the probability of the hypothesis tested but rather the probability of the data if some hypothesis (the null hypothesis) *were true* [p(D|H), the probability of the data given some hypothesis]. The goal of NHST is such that if the probability of the data given the null hypothesis is low enough, we might start thinking the data come from a world in which the null hypothesis is not true. Consequently, we reject the null hypothesis as a believable description of the population, and decide to believe an alternative explanation of events. Unfortunately, many researchers make the mistake of thinking that a failure to reject the null hypothesis has provided a probability that the null hypothesis is true – researchers may say: “a failure to reject the null hypothesis means that my groups are equal within some specified probability” – however, this is a conditional hypothesis that NHST is *not* testing.

Another misunderstood issue is interpreting the observed p-value in a valid way and choosing a corresponding cut-off value for the observed p-value. For some researchers, there is a rigid adherence to p = .05 as a cutoff point for significance (or some other e.g. p=.01). In other words, if the probability of the data under the null hypothesis is .045, these researchers will conclude to reject the null hypothesis. However, if the probability value is .055 (slightly above the cut-off), many researchers may not even discuss the result, or at best give it lower class status of significance (i.e. “marginal significance”). However, the decision whether to accept or reject the null hypothesis is inherently a subjective one, despite many interpretations to the contrary. To conclude that a result is “marginally” or “highly” significant is nonsensical. A statistical result is or isn’t statistically significant depending on the researcher’s point of view, and *regardless of the p-value obtained*. But what is exactly does this p-value represent?

P-Values and Error Rates

As mentioned previously, the observed p-value doesn’t represent the probability of the null hypothesis. Furthermore, the p-value also doesn’t tell us about the likelihood of any alternative hypothesis. Even if we had a specific alternative hypothesis, the p-value obtained with NHST only deals with the null hypothesis distribution of values (and a hypothetical one at that – e.g. can we really obtain a random sample from the population of all kids with ADHD?).

Historically, there have been at least two ways to approach “statistical significance”. In much of the social sciences, these two approaches are blended together in an almost incoherent fashion – and this hybrid has been promulgated in methodology texts. Fisher, a developer of the NHST methodology, even seemed to change his mind at one point as to how to interpret a NHST p-value. Fisher’s stance was that the observed p-value in NHST reflected our confidence in the null hypothesis. However, we already know is a problematic interpretation in the sense that the p-value is attached to the data (D), not to the hypothesis (H). Fisher also made no claims to an alternative hypothesis.

Neyman and Pearson, also developers of NHST, disagreed with Fisher’s approach. Neyman and Pearson’s approach was to specify an acceptable significance level before the experiment was conducted, and introduced the alpha cut-off (a) level, or Type I error rate (along with the concepts of: Type II error rate, power, and the alternative hypothesis). In the Neyman and Pearson approach: a researcher should, before data analysis, specify the probability of making a type I error (probability of incorrectly rejecting the null hypothesis when it is actually true). This specification will determine decisions about the design of the experiment (e.g. sample size for the experiment).

Thus, if I set the error rate at 5%, or a = .05, and I conduct the same experiment many times (all things being equal), and perform the corresponding analyses of the data, rejecting the null hypothesis 100 times, I will only be incorrect in doing so no more than five of those 100 replications. With this approach, it makes no difference whether the obtained p-value is .045 or .001, since we would make the same sort of decision, to reject the null, as long as our test statistic (e.g. observed t-value) falls beyond our specified cutoff point (critical value). In fact, the reporting of a specific p-value makes no sense in this approach - our statistic either makes the cut or it doesn’t based on our chosen alpha level.

The drawback with the Neyman-Pearson approach is that though we do have an idea as to a hypothetical long run situation of events, we are at a loss as to where our particular scenario resides within those hypothetically infinite number of random samples and analyses. In other words, we've rejected the null, but we'll never really know if this is the time that we've made the type I error.

What if the analysis does not allow a rejection of the null hypothesis- what does that mean? Fisher thought that it meant we weren't trying hard enough. Essentially, since we can't prove any hypothesis, only falsify it (e.g. in a Popperian sense), conclusions can't be drawn from a non-significant p-value. In other words, no matter how many white swans I see I can never prove that no black one exists, so if I don't see a black one I must keep looking. Despite these issues, Neyman and Pearson took a practical stance regarding this procedure. If we don't reach our cut-off, then according to the rules we've laid out, we act as though the null hypothesis were true (i.e. decide one course of action rather than another).

There are instances, however, where researchers using the N-P method will use Fisherian phrases like 'fail to reject the null'. In many journal articles and textbooks, researchers blend the two interpretations of NHST- the *epistemic* approach of Fisher: a procedure that tells us about the falsehood of a nil hypothesis, and the *behavioristic* Neyman-Pearson approach that allows for making decisions but does not really infer anything. These researchers will often specify an alpha level (cut-off level) and then interpret the p-value in the Fisherian sense. In fact, some researches will erroneously interpret the p-value as a kind of *effect size*, or strength of the finding (correlation, difference of means etc.) such that a p-value of .005 is representative of a stronger result than .035.

In Summary

The crux of the matter is that as researchers, sometimes little attention is paid to what the results mean practically, before moving on to a next set of analyses. Poor research design in the social sciences often make it difficult to detect important phenomenon from study to study (i.e. low sample size leading to low statistical power). Additionally, practices like rigid adherence to cut-off values despite inadequate sample size contribute to a lack of replicability of important phenomenon in the literature. Furthermore, editorial practices that tend only to publish statistically significant results (publication bias) have also led to spurious findings being reinforced in the literature as non-chance findings.

Poor methodological practice in the social science is a practice that encourages finding a significant result for data rather than approaching data with a thoughtful, problem-solving approach. Researchers that find themselves worrying about finding an observed p < .05, will find that their design will often ensure that such a result is found, and often be based on questions that are not all that interesting, with results that may be largely unenlightening. Confirmatory approaches combined with exploratory approaches (techniques that allow the data speak for itself), are flexible in the face of contradictory evidence, and assumes enough competency on the part of those who will be interested in the results to make decisions about the data for themselves.

What To Do?

A first step toward good statistical inference would be to recognize that the process of data analysis is more subjective than it was previously presumed to be. Researchers must make decisions every step of the way: interpreting previous results, formulating hypotheses, designing potential experiments, analyzing results, and deciding what is important to investigate further. As statistical methodology is a major tool that researchers use to study the data collected, researchers must be thoughtful in their approach and decision-making with regard to how they proceed at each stage of the analytical process. Decisions will have to be made on the part of the experimenter, and the researcher would be advised to be flexible, cautious, open-minded, and to use modern methods appropriate to the analysis situation.

Some Guidelines

1. The method of NHST described above and which is pervasive throughout much of the social sciences is not the only way to proceed with statistical inference. There are alternative methods like Bayesian and Likelihood inference.

2. Do not underestimate the initial analysis of data. Descriptive information is extremely important in understanding what information is in the data. It is in this initial stage that one can: find highly influential cases, detect errors in data entry that would otherwise bias the inferential statistics, discover other things to explore that hadn't been thought about previously, and better understand the results of the inferential analyses conducted later. One may even find that the initial type of analysis one wanted to carry out would perhaps not be the best choice, and that there may at the very least be a better way of going about it. In fact it might be the case that no further analysis is necessary.

3. When conducting NHST, report as much as possible- the more information the better. Report exact p-values, confidence intervals, effect sizes, any and everything you can think of that will help get your point across. Also do not be rigid in your interpretation of "significance". If it looks interesting to you, it probably would to someone else as well.

4. In the end, the p-value is of little importance on its own. More information about the result is to be gained from the reporting of an effect size, and there are various ones to choose from depending on your situation. Effect sizes give a measure of practical importance. As an example, I've shown my stats classes the "non-significant" result of a difference in grade average in which people were divided according to how often they attended class. The practical difference was 3 or 4 percentage points and a letter grade change, obviously important to them.

5. Be open-minded as to other interpretations. Your theory might be wrong. You should be more interested in finding out what is really going on than making the data conform to your expectations.

Significance testing is problematic, much more so than talked about here, and one is invited to look into some of the references provided below. Much of the problem seems to stem from a misunderstanding of the results. With a more careful approach and a basic understanding of the origins of the analyses we are conducting, NHST can provide much insight into the constructs social scientists concern themselves with.

Works Consulted

- Abelson, Robert. Statistics as Principled Argument. Mahwah, NJ:Erlbaum, 1995.
- Chatfield, C. (2002). Confessions of a Pragmatic Statistician. The Statistiction, 51, pp. 1-20.
- Cohen, J. (1994). The earth is round, p < .05. American Psychologist, 49, 997-1003.
- Cohen, J. (1990). Things I have Learned (So Far). American Psychologist, 45 (12), pp. 1304-1312.
- Gigerenzer, G. (1993). The Superego, The Ego and the Id in Statistical Reasoning. In Keren & Lewis (Eds.) Data Analysis in the Behavioral Sciences.
- Hubbard R. & Bayarri, M.J. (2003). Confusion Over Measures of Evidence (p's) Versus Errors (α's) in Classical Statistical Testing. The American Statistician. Volume: 57 Number: 3 Page: 171 – 178
- Oakes, M. 1986. Statistical Inference: A Commentary for the Social and Behavioral Sciences. Chichester, John Wiley & Sons.
- Rosnow, R, & Rosenthal, R. Effect Sizes for Experimenting Psychologists. Canadian Journal of Experimental Psychology, 57 (3), pp. 221-237.

Informative Web Links

http://www.cnr.colostate.edu/~anderson/thompson1.html